

国立国語研究所学術情報リポジトリ

『現代日本語書き言葉均衡コーパス』新聞サブコーパスの記事情報

メタデータ	言語: Japanese 出版者: 日本語学会 公開日: 2024-02-20 キーワード: 現代日本語書き言葉均衡コーパス, 日本語コーパス, 新聞記事, 情報付与 作成者: 加藤, 祥, 浅原, 正幸 メールアドレス: 所属:
URL	https://repository.ninjal.ac.jp/records/2000171

《資料・情報》

『現代日本語書き言葉均衡コーパス』
新聞サブコーパスの記事情報

加藤祥・浅原正幸

キーワード：日本語コーパス，新聞記事，情報付与

1. はじめに

『現代日本語書き言葉均衡コーパス（以降 BCCWJ）』の新聞サブコーパス（以降 PN）全サンプルについて、サンプル内に含まれる記事単位の分割を行うとともに、記事ごとの掲載紙面を確認し、記事分類情報や記事種別情報を付与した。これまでできなかった記事を単位とした分析や、特定の内容や種類の記事の抽出が可能となった。本稿では、新たに付与された新聞記事情報の設計と基礎統計データについて報告する。

2. 既存情報と本研究による付与情報

BCCWJ の PN には、ジャンル（全国紙・ブロック紙・地方紙の別）と朝刊夕刊の別、掲載年月日が付与されている。また、設計上、1つのサンプルには複数の記事が含まれている場合があり、複数記事は複数面にわたって掲載されている場合もある。さらに、掲載紙面や種別などに関する情報が付与されておらず、各記事別の扱いがしにくい。もちろん、BCCWJ を用いて新聞テキストの特徴を分析するにあたっては、報道や社説、コラムや連載小説などを含む新聞全体の実態を把握することが可能である。しかし、記事別の語彙調査や文体分析においては、新聞の「記事」に関する情報が必要な場合がある。本稿の追加情報によって、特定の内容や種類の記事を抽出することが可能となった。たとえば、記事分類と記事種別の情報を組み合わせて検索し、国内の事件報道や料理レシピの抽出ができる。連載小説のみを除外するようなことも可能となる。

3. 情報付与作業

BCCWJ の PN ファイル 1,473 サンプルを対象とした。作業者は4名で、BCCWJ 構築時にサンプリング対象となった実際の新聞紙面を手作業で確認した。サンプル内の記事境界情報（3.1）を付与し、各記事に内容情報（3.2）や種別情報（3.3）を付与した。

3.1 サンプル内の記事境界情報付与

記事境界の情報として、各記事の開始位置を付与し、サンプルごとの記事数も付した。また、記事の完結・未完結情報を付与した。

BCCWJ 内に付与されている記事タグ (article) を参照し、既存の記事タグの情報を利用した。よって、見出し部分のみが記事認定されている場合も 1 記事とした。しかし、たとえば「オリンピック」という大きな記事の中に競技別の小記事が確認されるような構造が見られる場合もあるため、記事種を付与するにあたって別記事の扱いが適当と考えられた際には、既存の記事タグ以下の細かな記事に分割した。このため、サンプル内の記事数は BCCWJ の記事タグ数とは一致しない。また、サンプル内の最後の記事は、取得サンプルの最後が可変長（文章のまとまりをもとに長さが決まる）の最後ではなく、固定長（サンプルの開始点を含む文の文頭から補助記号を除いた 1,000 字目のサンプルの終了点を含む文の文末までが収録されている）の最後であるとき、入力が途中で完結していない場合がある。そのため、サンプル内の各記事が記事末まで収録されているかを確認し、「記事の完結・未完結」情報を付与した。本作業の結果、PN の 1,473 サンプル中には 5,585 記事が認定された。このうち完結記事は 4,555 記事であった。

3.2 記事分類情報（記事内容情報）の付与

まず、各記事の掲載紙面情報として、掲載面の一般名称（総合・国際・社会など）の記載があれば取得した。掲載面に新聞固有の名称（「経済がわかる 企業がみえる」「いぶにんぐスベシャル」など）があれば参考として取得した。名称の記載がない場合には、「なし（一面）」「なし（ラテ欄）」などと（ ）で取得可能な情報を付した。また、各記事が国

表1 付与した記事分類

記事内容大分類		記事内容小分類									
0 総類	00 総類	01 皇室	02 言論報道	03 新聞	04 放送	05 出版	06 新メディア	07 大戦			
1 政治	10 政治	11 国会	12 選挙	13 政党	14 行政	15 地方	16 司法	17 財政	18 外交	19 防衛	
2 経済	20 経済	21 金融	22 貿易	23 商業	24 鉱工業	25 エネルギー	26 農林漁業	27 運輸交通	28 情報通信	29 土木建設	
3 労働	30 労働	31 雇用	32 労働条件	33 賃金	34 労働運動	35 労働組合	36 官公労組	37 民間労組	38 職業		
4 文化	40 文化	41 文学	42 教育	43 宗教	44 美術	45 芸能	46 演劇	47 映画	48 音楽		
5 科学	50 科学	51 地球	52 天文	53 宇宙	54 原子力	55 生物	56 医療	57 生理疫病	58 保健衛生	59 公害	
6 社会	60 社会	61 福祉	62 世代	63 家庭	64 生活	65 趣味娯楽	66 旅行観光	67 世相風俗	68 行事		
7 事件	70 事件	71 災害	72 火災	73 事故	74 交通事故	75 凶悪犯	76 人質犯	77 経済犯	78 公安事件	79 その他犯罪	
8 スポーツ	80 スポーツ	81 総合競技	82 陸上	83 水上	84 冬季	85 球技	86 野球	87 武道	88 諸競技	89 ギャンブル	
9 国際	90 国際	91 国際政治	92 国際軍事	93 平和運動	94 被爆	95 国際経済	96 国連				
見出し											

内・海外・国際（日本と他国に関わる）のいずれの内容であるのかを付与した。作業の結果、国内記事が4,530件、海外記事が736件、国際記事が284件（その他判別不能見出し等35件）である（短単位数については後述の表6を参照）。記事内容は、PNの含む全国紙（読売・朝日・毎日）の記事分類を参照し、公開される中で最も網羅で分類の偏りが少ない紀伊國屋書店『CD-HIASK』を基に、表1の分類を設定して付した。

作業者は各記事を読み、大分類（表1の0～9、不明瞭の場合は「見出し」）を選定した後、小分類を選定した。複数の内容が含まれる記事は、主と判断された分類が付与された。但し、記事内容に重複があるなど、どちらを主とするか判断揺れが生じる場合の作業基準は統一し（例：訴訟に関する記事で、「事件」か「司法」の揺れが生じるため、事件等を取り扱う内容であれば「事件」とし、「訴訟」「国会（代表質問）」などを備考欄に付す）、最終的に1名の作業者が整理を行った。また、地方紙において政治記事が全て「地方行政」となる可能性など、さらに下位の分類が必要となる場合には、「地方行政」であれば「訴訟」「財政」など、「諸競技」であれば「ボクシング」「ゴルフ」などの情報を備考欄に付した。大分類における記事数の分布を表2に示す（短単位数については後述の表6を参照）。

表2 記事内容（大分類）による記事分布（記事数）

大分類	社会	スポーツ	文化	政治	経済	事件	国際	総類	科学	労働	他	総計
記事数	1038	819	705	698	666	502	363	338	281	106	69	5585

3.3 記事種別情報の付与

記事の種別として、2種類の分類情報を付与した。紀伊國屋書店『CD-HIASK』の種別を援用した以下の分類項目を設定した。作業者は各記事を読んで情報を付与した。分類例と分類分布（記事種別1は表3、記事種別2は表4）を示す。これらの記事種別情報を付与した結果、典型的な新聞記事と考えられる報道記事（特殊な記事種別が付与されない記事）、評論や連載などの抽出が可能となった。

【記事種別1】

評論：硬めの囲み記事（社説や新聞社コラムを含む）

- （1）今年の国内政治での最大の焦点は夏の参院選だ。一向に出口が見えない長期の不況、社会全体をおおう閉塞状況、重なる将来への不安など前世紀終盤から引き継いだ課題への各党の対応を、国民はどう判断するか。二十一世紀最初の国政選挙の結果が大いに注目される。（サンプルID：PN1b_00002、以下同様に示す）

投書：囲み記事で「投書」とあるか、投書の引用が明示的な記事

- （2）中学の時に剣道部顧問が異動、名義だけの副顧問だった養護の先生が急きょ代役

になり、以来一年間、道場に。受けた助言といえば、「風邪をひいたら水分を取る」くらいですが、とにかく熱心。子どももいるのに、毎日ちゃんと練習終了までいてくれました。(PN1g_00004)

インタビュー：発言がそのまま記述されている記事（討論の場合は備考に付す）

- (3) —今の経済状態は「非常事態に当たる」と。◆そうです。日本の金融機関の不良債権問題を調べにきたアメリカ側が、その惨状に腰を抜かした。(PN1b_00004)

ハウツー：料理レシピや手芸指南等の記事

- (4) パン粉は牛乳で湿らせ、タマネギはみじん切りにしてバターであめ色にいためる。豚ひき肉はどこ部位でもいいが、脂身が多い方がうまみが出る。赤身の部位の時は、豚バラ肉やベーコンを刻んで加えてもいい。(PN3c_00011)

催し案内：日時等が紹介されている記事（イベント情報・テレビ番組等含む）

- (5) 自分を知る家族劇イエロードロップスの「スマイリー」世田谷区北沢「劇」小劇場来月3～8日「紙芝居屋」で生計を立てる父娘。(中略)企画集団「イエロードロップス」(佐東諒一さん主宰)が7月3～8日、世田谷区北沢2の「劇」小劇場で、演劇「スマイリー」を公演する。(PN1b_00016)

人：人物紹介、評伝などの記事（囲み記事であってもインタビューではない場合）

- (6) ソプラノ歌手笹田 和子マイスターのような誠実さも ささだ・かずこ 大正十年、大阪府生まれ。昭和十七年、東京音楽学校（現・東京芸大）を卒業後、藤原歌劇団によるワーグナー「ローエングリン」（日本初演）のエルザ役でデビュー。(PN1d_00026)

色：家庭、文化、娯楽などの囲み記事（囲碁将棋の枠なども含む）

- (7) 第五十六期 本因坊戦挑戦者決定リーグ九段 結城 聡 (1勝2敗)先番 八段 山田規三生 (3敗)【第十三局の5】百三―百二十五(前回の最終手は白9ニ)(PN1b_00006)

- (8) 慢性腰痛の人には、腹筋と背筋のバランスの不良が多くみられます。腹筋が弱くなって緩み、背筋の緊張が強くなると、腰椎が後ろに反って腰痛を起こしやすい姿勢になるのです。腹筋とともに背筋力も弱くなり、腰椎を支える力が全体的に低下するということも少なくありません。(PN1h_00022)

人事：人事異動等に関する記事

- (9) 石播社長に伊藤専務昇格 石川島播磨重工業は二十日までに、伊藤源嗣専務が次期社長に昇格するトップ人事を固めた。武井俊文社長は相談役に退く見通し。
(PN1j_00021)

募集：読者の応募を求める記事

- (10) メガネ無料券 日本オプティカルは、(中略)「ジャスト3 プライスマメガネ販売」のスタートを記念して、「メガネ無料券」(五千円相当)を5人にプレゼント。はがきに〒住所・氏名・年齢・職業・電話番号を明記。(PN1f_00001)

賞：受賞に関する記事

- (11) 緒方貞子さんに独が大功労勲章 緒方貞子・前国連難民高等弁務官は二十三日、ドイツ政府から「大功労十字星章付大綬章」を贈られた＝写真。(PN1a_00009)

表3 記事種別1による記事分布(記事数)

記事種別1	評論	投書	インタビュー	ハウツー	催し	人	色	人事	募集	賞	一般
記事数	366	33	90	30	427	240	286	7	36	8	4062

【記事種別2】「テキスト(一般)」のほか「略歴」と以下がある。**連載：連載番号または「上中下」「前後」等が確認できた記事(小説も含む)**

- (12) アメリカ星条旗のもとで■上 宗教と建国精神「開かれた社会」続ける分岐点 一枚岩ではない戦争支持(中略)九月十一日以来、アメリカのあらゆる所で、非公式の国歌と呼ばれる「ゴッド・ブレス・アメリカ」(神よ、アメリカを守り給え)が歌われ、全国津々浦々に星条旗がひるがえった。(PN1a_00025)

解説：報道記事の説明部分、事件事象等の囲み説明・催しの囲み説明等

- (13) 編集長になる佐藤純一さんは、「ソクラテスがアゴラ(ギリシャ語で市場)で対話したように、人々が集まって話せる街角のような雑誌にしたい」と抱負を話す。「人間会議」は、「宣伝会議」「ブレーン」など歴史のある雑誌を刊行している出版社「宣伝会議」(東京都港区)の新雑誌。(PN1d_00017)

日誌：当該日付の出来事紹介、記者日記等

- (14) きょうのこよみ＝二十三日＝(中略)▽千九百九十六(平成8)年 オーストラリアのブリスベンで開かれた第6回ラムサール条約締約国会議で、新潟市の佐潟が、国内では十番目の登録湿地として認定された。(PN1i_00013)

用語：用語解説・用語説明

- (15) ことば 炭素税構想 地球温暖化の原因となる二酸化炭素の排出量削減のため、石炭や石油などの化石燃料に、炭素の排出量に応じて課税する炭素税を導入する構想。(PN1k_00021)

データ：日時のみをはじめ数値情報・式辞発表・判決などの全文掲載等

(データ表等は BCCWJ ではサンプリング対象外のため該当なし)

- (16) 漁況マサバ立て縄・多針釣り漁＝ 大正礁および室戸岬周辺の立て縄・多針釣りは、甲浦へ十二隻ゴマサバ 0.2 トン（一日 1 隻当たり十七キロ）の水揚げ。(PN4m_00003)

死亡：死亡記事（訃報、おくやみ・解説・略歴等の囲み記事形式も含む）

- (17) 勝村 龍雄さん（かつむら・たつお＝元大阪曹達〈現ダイソー〉社長）二十四日、老衰で死去、九十歳。通夜と葬儀は近親者のみで行った。ダイソーなど主催の「お別れの会」は十二月十八日午前十一時三十分から大阪市北区中之島 5 の 3 の六十八のリーガロイヤルホテル桐の間で。(PN2a_00027)

表 4 記事種別 2 による記事分布（記事数）

記事種別 2	連載	解説	テキスト	略歴	日誌	用語	データ	死亡	その他
記事数	357	768	4320	5	42	7	26	22	38

なお、記事種別 1 及び記事種別 2 は各記事に付与されているため、連載される評論（図 1 左）や解説が囲み記事となっている場合（図 1 右）のような組み合わせが生じる。



図 1 左：評論×連載の例 (PN1c_00004)
読売新聞 朝刊 2001/8/24



右：色×解説の例 (PN5a_00003)
朝日新聞 朝刊 2005/7/8

また、記事種別の記事内容別の分布も確認できる。**表 5** に記事種別 1 の記事内容別分布を示す。囲み記事（色）が「社会」記事に多い、催し案内が「総類」「文化」記事に多いなどの特徴が確認できる。

表 5 記事種別 1 の記事内容別分布（記事数）

記事種別 1 \ 内容	スポーツ	科学	経済	国際	事件	社会	政治	総類	文化	労働	総計
評論	27	33	60	42	26	53	71	37	110	10	476
催し案内	10	9	9	5	4	93	1	179	133	2	449
色	3	17	5		3	229	2	3	34	1	298
人	69	9	21	5		48	12	16	65	18	266
インタビュー	20	10	18	11	12	24	15	6	31	4	151
投書		1	1	1	1	36	3		8		53
募集			2			17	1	5	4	2	35
ハウツー	1		4			20			5		30
人事	2		7				13		2	2	26
賞		1	1			1		1	9		13

4. 集計例

4.1 短単位分布

記事内容や記事種別に基づく短単位数の基礎統計情報を示す。**表 6** に、記事内容（大分類）の短単位分布を国内外情報と組み合わせて集計した例を示す。但し、「見出し」は国内外情報の分類ができなかった例がある（64 短単位）。なお、表 2 の記事数分布と表 6 の短単位数分布では、頻度順位が異なる。たとえば「スポーツ」記事は 1 記事あた

表 6 記事内容（大分類）による短単位の国内外記事分布（短単位）

記事内容（大分類）	短単位	海外	国際	国内
社会	305165	10986	4487	289692
政治	223753	30420	5115	188218
文化	220310	8387	902	211021
スポーツ	215547	22748	12107	180692
経済	166000	13234	4252	148514
事件	130026	18149	2153	109724
総類	115730	3570	3252	108908
国際	108973	41885	32428	34660
科学	90853	6954	0	83899
労働	29634	1426	0	28208
見出	226	28	0	134
総計	1606217	157787	64696	1383670

りの短単位数が少ないが、「政治」記事は多い傾向が確認できる。また、国内の事象のみが扱われている短単位数の割合が86%（1,383,670 短単位）を占めるが、「国際」記事の短単位数は、海外または国際、国内分布においてほとんど差がないなどの特徴が見られる。

4.2 BCCWJ-WLSP との重ね合わせによる記事内容別語義分布

分類語彙表番号を付与した BCCWJ（BCCWJ-WLSP, 加藤ら 2019）と重ね合わせることにより、意味的な観点による集計が可能である。BCCWJ-WLSP の PN は 117,543 短単位である。表 7 では、記事内容（大分類）の短単位分布を分類語彙表番号の中項目において集計した。ジャンル別の意味的な分布が確認できる。分類番号によるさらに詳細な集計も可能であるため、記事内容による意味特徴の調査に活用可能性がある。

表 7 分類語彙表番号（中項目）と記事内容（大分類）の短単位分布（中項目上位 20 位）

大分類 中項目	スポーツ	科学	経済	国際	事件	社会	政治	総類	文化	労働	総計
1.19_量	1659	482	2002	552	1109	1313	2186	486	1836	82	11707
1.25_公私	363	160	479	417	379	385	549	197	566	17	3512
1.30_心	272	140	459	180	254	439	925	119	470	46	3304
2.34_行為	311	101	374	210	265	364	633	126	391	37	2812
1.16_時間	390	83	387	131	225	368	450	153	478	29	2694
2.15_作用	293	118	256	94	148	365	313	101	397	23	2108
2.12_存在	241	84	293	108	191	339	375	73	325	11	2040
1.15_作用	228	127	367	133	152	249	453	63	217	9	1998
1.31_言語	93	40	207	193	151	156	295	79	468	23	1705
1.37_経済	65	34	644	57	86	276	331	37	84	8	1622
1.27_機関	184	39	221	110	165	85	556	50	160	15	1585
1.20_人間	134	43	118	37	268	218	265	114	341	37	1575
1.24_成員	229	23	131	104	126	126	362	85	230	12	1428
1.17_空間	192	65	198	68	147	190	253	44	191	10	1358
1.26_社会	114	44	266	54	116	265	118	61	308	6	1352
1.35_交ワリ	333	16	122	149	61	122	337	64	132	16	1352
1.36_待遇	46	10	156	92	120	209	501	35	148	23	1340
1.10_事柄	135	41	165	72	136	173	278	41	212	14	1267
1.38_事業	62	54	321	54	98	207	197	60	175	6	1234
付属語	6622	1982	6602	3034	4313	7463	9417	2666	9553	529	52181
総計	14230	4615	16124	6890	10108	16383	22024	5865	20124	1180	117543

5. おわりに

BCCWJ に新たに付与された新聞記事情報について報告した。本稿のデータにより、ジャンル（記事内容）の絞り込みや、典型的な新聞記事（報道記事）の抽出、連載小説の影響の確認などに活用が可能であると期待される。本稿のデータは国語研に依頼し、リポジトリにて公開予定であり、広範な利用が可能となる。

引用参考文献一覧

- 『CD－HIASK〈2004〉朝日新聞記事データベース』紀伊國屋書店
- 丸山岳彦，山崎誠，柏野和佳子，佐野大樹，秋元祐哉，稲益佐知子，田中弥生，大矢内夢子，2011.『『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装』
- 加藤祥，浅原正幸，山崎誠，2019.「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」『日本語の研究』vol. 15, No. 2, pp. 134-141.
- 国立国語研究所（編），2004.『分類語彙表増補改訂版データベース』

——かとう さち 目白大学専任講師——

——あさはら まさゆき 国立国語研究所教授——

（2022年9月23日 第1稿受理）

（2023年4月28日 最終稿受理）